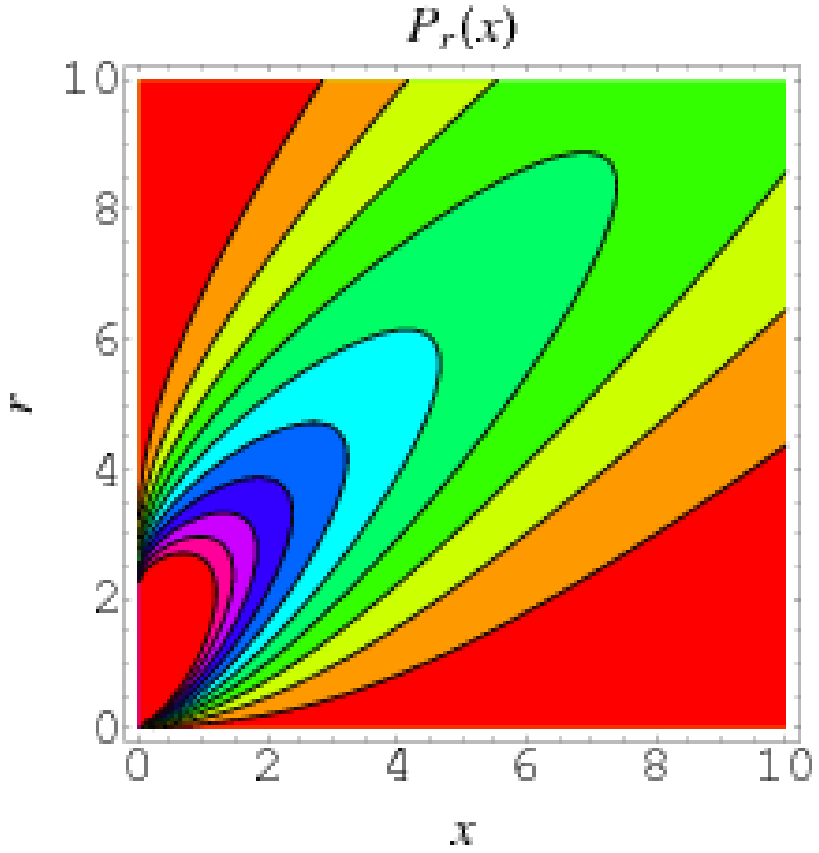
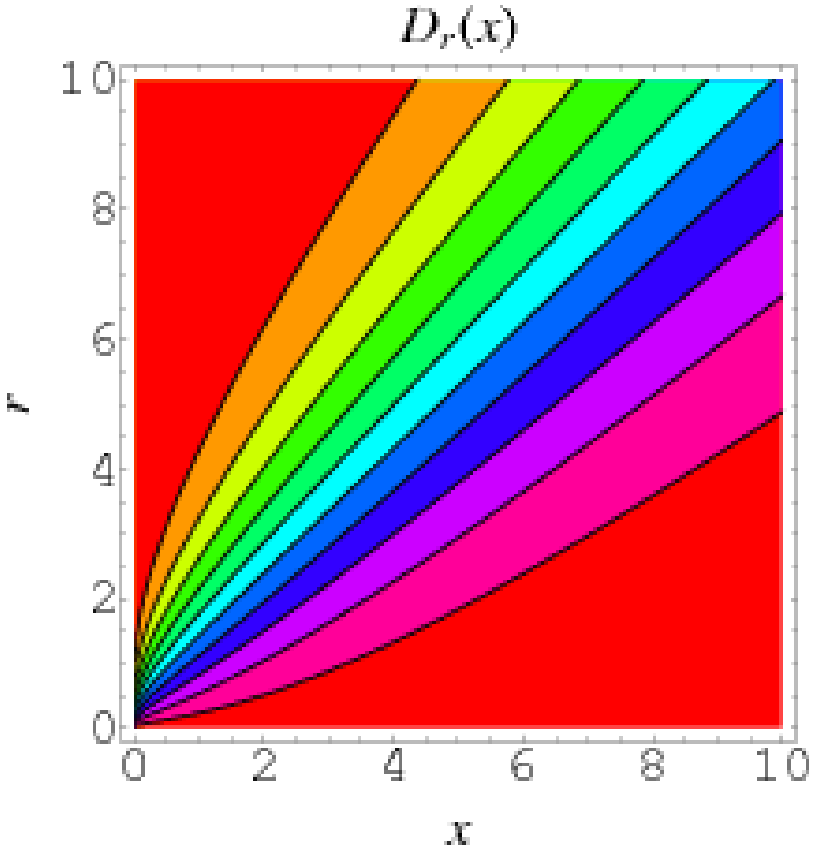
More stats...
Outliers, and Chi²





http://mathworld.wolfram.com/Chi-SquaredDistribution.html

THE FOLLOWING **PREVIEW** HAS BEEN APPROVED FOR

ALL AUDIENCES

BY THE MOTION PICTURE ASSOCIATION OF AMERICA, INC.

www.filmratings.com

www.mpaa.org

- Inquiry 1 written and oral reports are due in lab T 9/20 or W 9/21
- Also need to start putting together your group for inquiry 2... 3-5 people/group
- Homeworks coming soon
- Online evaluation
- TA office hours calendar online
- Stream sort
- Statistics quiz key posted

Outliers…

2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 5, 6, 7, 121, 130

Median = 4

Mean = 18

Outliers: When is data invalid?

Outliers: When is data invalid?

Not simply when you want it to be.

Outliers: When is data invalid?

Not simply when you want it to be.

Dixon's Q test can determine if a value is statistically an outlier.

Dixon's Q test can determine if a value is statistically an outlier.

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)|$$

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)| = 0.742$$

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

Dixon's Q test can determine if a value is statistically an outlier.

Example: results from a blood test…
789, 700, 772, 766, 777

$$Q = |(700 - 766)| \div |(789 - 700)| = 0.742$$ **So?**

$$Q = \frac{|(\text{suspect value} - \text{nearest value})|}{|(\text{largest value} - \text{smallest value})|}$$

You need the critical values for Q table:

Sample #	Q critical value
3	0.970
4	0.831
5	0.717

If Q calc > Q crit
rejected

From: E.P. King, J. Am. Statist. Assoc. 48: 531 (1958)

You need the critical values for Q table:

Sample #	Q critical value
3	**0.970**
4	**0.831**
5	**0.717**

If Q calc > Q crit than the outlier can be rejected

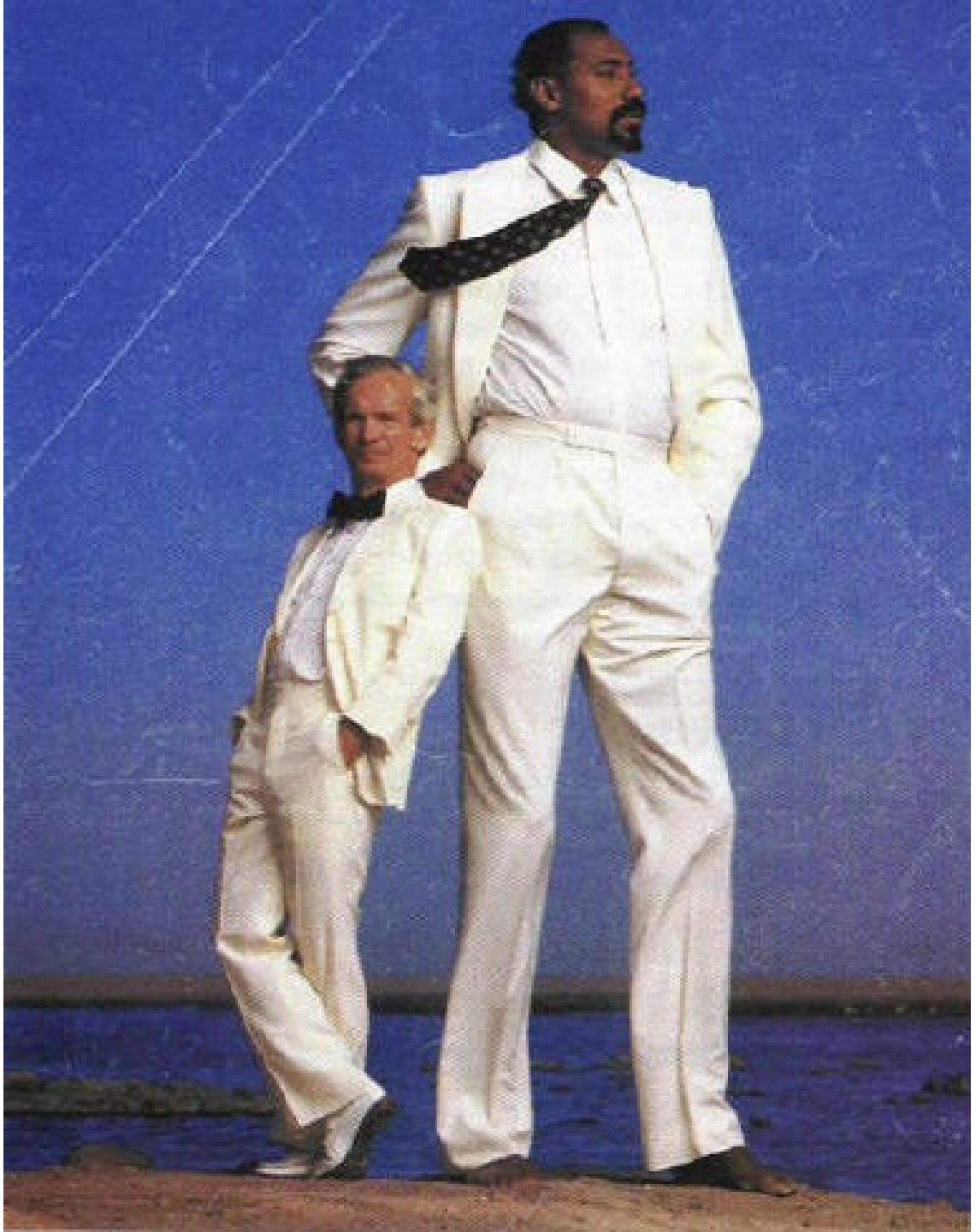
Q calc = 0.742

Q crit = 0.717

= rejection

From: E.P. King, J. Am. Statist. Assoc. 48: 531 (1958)

What can outliers tell us?



If you made a mistake, you <u>should have</u> already accounted for that.

Outliers can lead to important and fascinating discoveries.

Transposons "jumping genes" were discovered because they did not fit known modes of inheritance.

The Chi Square Test

- A statistical method used to determine goodness of fit
  - Goodness of fit refers to how close the observed data are to those predicted from a hypothesis

The Chi Square Test

- The general formula is

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

- where
  - O = observed data in each category
  - E = observed data in each category based on the experimenter's hypothesis
  - Σ = Sum of the calculations for each category

Two flies with different traits are bred together

- Out of 352 offspring
  - 193 straight wings, gray bodies
  - 69 straight wings, ebony bodies
  - 64 curved wings, gray bodies
  - 26 curved wings, ebony bodies

According to our hypothesis, there should be a 9:3:3:1 ratio of fly offspring

Phenotype	Expected probability	Expected number
straight wings, gray bodies	9/16	9/16 X 352 = 198
straight wings, ebony bodies	3/16	3/16 X 352 = 66
curved wings, gray bodies	3/16	3/16 X 352 = 66
curved wings, ebony bodies	1/16	1/16 X 352 = 22

Apply the chi^2 formula

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$$

$$\chi^2 = \frac{(193 - 198)^2}{198} + \frac{(69 - 66)^2}{66} + \frac{(64 - 66)^2}{66} + \frac{(26 - 22)^2}{22}$$

$$\chi^2 = 0.13 + 0.14 + 0.06 + 0.73$$

$$\chi^2 = 1.06$$

- Interpret the chi square value
  - The calculated chi square value can be used to obtain probabilities, or P values, from a chi square table
    - These probabilities allow us to determine the likelihood that the observed deviations are due to random chance alone

  - If the chi square value results in a probability that is less than 0.05 (ie: less than 5%)
    - The hypothesis is rejected

- Interpret the chi square value
  - Before we can use the chi square table, we have to determine the degrees of freedom (*df*)
    - The *df* is a measure of the number of categories that are independent of each other
    - $df = n - 1$
      - where n = total number of categories
    - In our experiment, there are four categories
      - Therefore, df = 4 – 1 = 3

TABLE 2.1

Chi Square Values and Probability

						Null Hypothesis rejected	
Degrees of Freedom	$P = 0.99$	0.95	0.80	0.50	0.20	0.05	0.01
1.	0.000157	0.00393	0.0642	0.455	1.642	3.841	6.635
2.	0.020	0.103	0.446	1.386	3.219	5.991	9.210
3.	0.115	0.352	1.005	1.06 2.366	4.642	7.815	11.345
4.	0.297	0.711	1.649	3.357	5.989	9.488	13.277
5.	0.554	1.145	2.343	4.351	7.289	11.070	15.086
6.	0.872	1.635	3.070	5.348	8.558	12.592	16.812
7.	1.239	2.167	3.822	6.346	9.803	14.067	18.475
8.	1.646	2.733	4.594	7.344	11.030	15.507	20.090
9.	2.088	3.325	5.380	8.343	12.242	16.919	21.666
10.	2.558	3.940	6.179	9.342	13.442	18.307	23.209
15.	5.229	7.261	10.307	14.339	19.311	24.996	30.578
20.	8.260	10.851	14.578	19.337	25.038	31.410	37.566
25.	11.524	14.611	18.940	24.337	30.675	37.652	44.314
30.	14.953	18.493	23.364	29.336	36.250	43.773	50.892

From Fisher, R. A., and Yates, F. (1943) *Statistical Tables for Biological, Agricultural, and Medical Research.* Oliver and Boyd, London.

- With *df* = *3*, the chi square value of 1.06 is slightly greater than 1.005 (which corresponds to P= 0.80)

- A P = 0.80 means that values equal to or greater than 1.005 are expected to occur 80% of the time based on random chance alone

- Therefore, it is quite probable that the deviations between the observed and expected values in this experiment can be explained by random chance

Spreadsheet applications will compute chi^2

Is the male:female ratio in the CNS different from the general population?

observed	expected	
40	50	male
60	50	female
Chi-sq =	0.0455	

3rd Thursday at Blanton Art Museum

(http://blantonmuseum.org/calendar_events/details/third_thursday21)

Rachel Harrison
Buddha with Wall, 2004

- Inquiry 1 written and oral reports are due in lab on T 9/20 or W 9/21
- Also need to start putting together your group for inquiry 2... 3-5 people/group
- Homeworks coming soon
- Online evaluation
- TA office hours calendar online
- Stream sort
- Statistics quiz key posted